Retarded kernels for longitudinal survival analysis and dynamic prediction

—— Supplementary Material ——

Annabel L Davies[1], Anthony C C Coolen[23] and Tobias Galla[14]

S1 Mathematical details

S1.1 Maximum Likelihood Inference

As in standard Cox survival analysis, we use maximum likelihood inference to determine the most plausible values of the model parameters for our models, based on the observed data. We write θ for the full set of parameters, this includes the base hazard rate $h_0(t)$. That is, $\theta = \{h_0(t), a_\mu, \tau_\mu; \mu = 1, \ldots, p\}$ for both models A and B. The optimal parameters are those for which the data likelihood $\mathcal{P}(\mathcal{D}|\theta)$ is maximised. We use the primary event indicator $\delta_i = I(T_i^* \leq C_i) \in \{0, 1\}$, where the indicator function $I(A)$ is defined as $I(A) = 1$ if A holds, and $I(A) = 0$ otherwise. The data likelihood for censored data is then

$$\mathcal{P}(\mathcal{D}|\theta) = \prod_{i=1}^{N} h(T_i|\theta, \mathcal{Z}^i_{[0,s_i]})^{\delta_i} S(T_i|\theta, \mathcal{Z}^i_{[0,s_i]}), \quad S(T_i|\theta, \mathcal{Z}^i_{[0,s_i]}) = \mathrm{e}^{-\int_0^{T_i} h(t|\theta, \mathcal{Z}^i_{[0,s_i]})\mathrm{d}t}, \tag{S1}$$

where, in this section, we refer only to the retarded kernel model and therefore omit the superscript 'RK' from the hazard for clarity. Maximizing $\mathcal{P}(\mathcal{D}|\theta)$ is equivalent to minimizing the negative log likelihood, i.e. $\hat{\theta}_{\mathrm{ML}} = \mathrm{argmin}_\theta\, \Omega_{\mathrm{ML}}(\theta)$, with

$$\begin{aligned}\Omega_{\mathrm{ML}}(\theta) &= -\log \prod_{i=1}^{N} h(T_i|\theta, \mathcal{Z}^i_{[0,s_i]})^{\delta_i} S(T_i|\theta, \mathcal{Z}^i_{[0,s_i]}) \\ &= -\sum_{i=1}^{N} \delta_i \log h(T_i|\theta, \mathcal{Z}^i_{[0,s_i]}) + \sum_{i=1}^{N} \int_0^{T_i} h(t|\theta, \mathcal{Z}^i_{[0,s_i]})\mathrm{d}t. \end{aligned} \tag{S2}$$

The hazard rate for the retarded kernel approach is

$$h(t|\theta, \mathcal{Z}^i_{[0,s_i]}) = h_0(t) \exp\left\{ \sum_{\mu=1}^{p} \int_0^{\min(s_i,t)} \beta_\mu(t, t', s_i) z^i_\mu(t')\mathrm{d}t' \right\}. \tag{S3}$$

Substituting this into $\Omega_{\mathrm{ML}}(\theta)$ yields

$$\begin{aligned}\Omega_{\mathrm{ML}}(\theta) = -\sum_{i=1}^{N} \delta_i \log h_0(T_i) - \sum_{i=1}^{N} \delta_i \int_0^{s_i} \sum_\mu \beta_\mu(T_i, t', s_i) z^i_\mu(t')\mathrm{d}t' \\ + \sum_{i=1}^{N} \int_0^{T_i} h_0(t') \mathrm{e}^{\int_0^{\min(s_i,t')} \mathrm{d}t'' \sum_\mu \beta_\mu(t',t'',s_i) z^i_\mu(t'')} \mathrm{d}t' \,, \end{aligned} \tag{S4}$$

[1]Department of Physics and Astronomy, University of Manchester, UK
[2]Department of Biophysics, Radboud University, The Netherlands
[3]Saddle Point Science Ltd, UK
[4]Instituto de Física Interdisciplinar y Sistemas Complejos, IFISC (CSIC-UIB), Campus Universitat Illes Balears, Palma de Mallorca, Spain

Corresponding author:
Annabel L Davies, Theoretical Physics, Department of Physics and Astronomy, School of Natural Sciences, The University of Manchester, Manchester M13 9PL, United Kingdom
Email: annabel.davies@postgrad.manchester.ac.uk

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

where we have used the fact that $s_i \leq T_i$. For simplicity we have specified the hazard in Equation (S4) without fixed (or baseline) covariates. To include these explicitly (if desired), one can simply add the term $\sum_\nu \gamma_\nu \zeta_\nu^i$ to the exponent of the hazard function.

Extremisation of Equation (S4) functionally over $h_0(t)$ gives the maximum likelihood estimator of the base hazard rate, given $\beta_\mu(t,t',s)$,

$$\hat{h}_0(t) = \frac{\sum_{i=1}^N \delta_i \delta(t-T_i)}{\sum_{i=1}^N I(t\in[0,T_i]) \mathrm{e}^{\int_0^{\min(s_i,t)} \sum_\mu \beta_\mu(t,t',s_i) z_\mu^i(t')\mathrm{d}t'}}, \tag{S5}$$

as quoted in Equation (21) in the main paper. Equation (S5) is the analogue of the standard Breslow estimator.[1] Inserting this expression back into Equation (S4) leaves the following function to be extremized over the remaining model parameters $\{a_\mu, \tau_\mu\}$ in the kernels $\beta_\mu(t,t',s)$ (where we denote all terms that do not contain $\{a_\mu,\tau_\mu\}$ simply as 'constant'):

$$\begin{aligned}
\Omega_{\mathrm{ML}}[\{a_\mu,\tau_\mu\}] &= -\sum_{i=1}^N \delta_i \log\left(\frac{\sum_{k=1}^N \delta_k \delta(T_i-T_k)}{\sum_{j=1}^N I(T_i\in[0,T_j]) \mathrm{e}^{\sum_\mu \int_0^{\min(s_j,T_i)} \beta_\mu(T_i,t',s_j) z_\mu^j(t')\mathrm{d}t'}}\right) \\
&\quad -\sum_{i=1}^N \delta_i \sum_\mu \int_0^{s_i} \beta_\mu(T_i,t',s_i) z_\mu^i(t')\mathrm{d}t' \\
&\quad +\sum_{i=1}^N \int_0^{T_i} \left(\frac{\sum_{k=1}^N \delta_k \delta(t'-T_k)}{\sum_{j=1}^N I(t'\in[0,T_j]) \mathrm{e}^{\sum_\mu \int_0^{\min(s_j,t')} \beta_\mu(t',t'',s_j) z_\mu^j(t'')\mathrm{d}t''}}\right) \mathrm{e}^{\sum_\mu \int_0^{\min(s_i,t')} \beta_\mu(t',t'',s_i) z_\mu^i(t'')\mathrm{d}t''} \mathrm{d}t' \\
&= \sum_{i=1}^N \delta_i \left\{ \log\left(\sum_{j=1}^N I(T_i\in[0,T_j]) \mathrm{e}^{\sum_\mu \int_0^{\min(s_j,T_i)} \beta_\mu(T_i,t',s_j) z_\mu^j(t')\mathrm{d}t'}\right) - \sum_\mu \int_0^{s_i} \beta_\mu(T_i,t',s_i) z_\mu^i(t')\mathrm{d}t' \right\} \\
&\quad +\sum_{k=1}^N \delta_k \left(\frac{\sum_{i=1}^N I(T_k\in[0,T_i]) \mathrm{e}^{\sum_\mu \int_0^{\min(s_i,T_k)} \beta_\mu(T_k,t'',s_i) z_\mu^i(t'')\mathrm{d}t''}}{\sum_{j=1}^N I(T_k\in[0,T_j]) \mathrm{e}^{\sum_\mu \int_0^{\min(s_j,T_k)} \beta_\mu(T_k,t'',s_j) z_\mu^j(t'')\mathrm{d}t''}}\right) + \text{constant} \\
&= \sum_{i=1}^N \delta_i \left\{ \log\left(\sum_{j=1}^N I(T_i\in[0,T_j]) \mathrm{e}^{\sum_\mu \int_0^{\min(s_j,T_i)} \beta_\mu(T_i,t',s_j) z_\mu^j(t')\mathrm{d}t'}\right) - \sum_\mu \int_0^{s_i} \beta_\mu(T_i,t',s_i) z_\mu^i(t')\mathrm{d}t' \right\} \\
&\quad + \text{constant}.
\end{aligned} \tag{S6}$$

This is the formula quoted in Equation (22) in the main paper. Minimisation of Equation (S6) with respect to the remaining model parameters $\{a_\mu,\tau_\mu;\mu=1\ldots p\}$ must be performed numerically. Finally, if we define the N^2 integrals

$$\mathcal{I}_{ij}[\{a_\mu,\tau_\mu\}] = \int_0^{\min(s_j,T_i)} \sum_{\mu=1}^p \beta_\mu(T_i,t',s_j) z_\mu^j(t')\mathrm{d}t', \tag{S7}$$

then we can re-write expression (S6) as

$$\Omega_{\mathrm{ML}}[\{a_\mu,\tau_\mu\}] = \sum_{i=1}^N \delta_i \log\left(\sum_{j=1}^N I(T_i\in[0,T_j])\, \mathrm{e}^{\mathcal{I}_{ij}[\{a_\mu,\tau_\mu\}]-\mathcal{I}_{ii}[\{a_\mu,\tau_\mu\}]}\right) + \text{constant}. \tag{S8}$$

S1.2 Survival probability

We recall from Equation (23) in the main paper that the estimated probability that subject i has not experienced an event by time $u > s_i$ conditional on their survival to s_i and on their covariate values $\mathcal{Z}^i$ up to that time is given by

$$\hat{\pi}^{\mathrm{RK}}(u|\mathcal{Z}^i_{[0,s_i]}, s_i) = \exp\left\{-\int_{s_i}^u \hat{h}(t'|\mathcal{Z}^i_{[0,s_i]})\mathrm{d}t'\right\}. \tag{S9}$$

Substituting into this equation the base hazard estimator in Equation (S5), in combination with with Equation (S3), yields

$$\begin{aligned}
\hat{\pi}^{\mathrm{RK}}(u|\mathcal{Z}^i_{[0,s_i]}, s_i) &= \exp\left\{ -\int_{s_i}^{u} \frac{\mathrm{e}^{\sum_{\mu=1}^{p} \int_0^{\min(s_i,t')} \hat{\beta}_\mu(t',t'',s_i) z^i_\mu(t'')\mathrm{d}t''} \sum_{j=1}^{N} \delta_j \delta(t'-T_j)}{\sum_{k=1}^{N} I(t' \in [0,T_k]) \mathrm{e}^{\int_0^{\min(s_k,t')} \mathrm{d}t'' \sum_\mu \hat{\beta}_\mu(t',t'',s_k) z^k_\mu(t'')}} \mathrm{d}t' \right\} \\
&= \exp\left\{ -\sum_{j=1}^{N} \delta_j \, I(T_j \in [s_i,u]) \frac{\mathrm{e}^{\sum_{\mu=1}^{p} \int_0^{\min(s_i,T_j)} \hat{\beta}_\mu(T_j,t'',s_i) z^i_\mu(t'')\mathrm{d}t''}}{\sum_{k=1}^{N} I(T_j \in [0,T_k]) \mathrm{e}^{\int_0^{\min(s_k,T_j)} \mathrm{d}t'' \sum_\mu \hat{\beta}_\mu(T_j,t'',s_k) z^k_\mu(t'')}} \right\} \\
&= \exp\left\{ -\sum_{j=1}^{N} \delta_j \, I(T_j \in [s_i,u]) \frac{\mathrm{e}^{\sum_{\mu=1}^{p} \int_0^{s_i} \hat{\beta}_\mu(T_j,t'',s_i) z^i_\mu(t'')\mathrm{d}t''}}{\sum_{k=1}^{N} I(T_j \in [0,T_k]) \mathrm{e}^{\int_0^{\min(s_k,T_j)} \mathrm{d}t'' \sum_\mu \hat{\beta}_\mu(T_j,t'',s_k) z^k_\mu(t'')}} \right\},
\end{aligned} \tag{S10}$$

where in the last line we replaced $\min(s_i, T_j) = s_i$, which holds by virtue of the factor $I(T_j \in [s_i, u])$. We have also used the notation $\hat{\beta}_\mu(t, t', s)$ to indicate the association kernel obtained from the ML estimators of the parameters $\{a_\mu, \tau_\mu\}$. Using the integral defined in Equation (S7) we can re-write Equation (S10) as

$$\hat{\pi}^{\mathrm{RK}}(u|\mathcal{Z}^i_{[0,s_i]}, s_i) = \exp\left\{ -\sum_{j=1}^{N} \delta_j \, I(T_j \in [s_i,u]) \frac{\mathrm{e}^{\mathcal{I}_{ji}[\{\hat{a}_\mu \hat{\tau}_\mu\}]}}{\sum_{k=1}^{N} I(T_j \in [0,T_k]) \mathrm{e}^{\mathcal{I}_{jk}[\{\hat{a}_\mu \hat{\tau}_\mu\}]}} \right\}, \tag{S11}$$

where we recall that i labels the individual for whom we are making predictions, while the sums over j and k refer to individuals in the data set used for inference.

S1.3 Step function interpolation

In the main paper we use staircase functions as a straightforward method to interpolate between discrete measurements of the covariates. We take a 'nearest neighbour' approach, that is we set $z^i_\mu(t) = z^i_\mu(t_{i\ell})$ where $t_{i\ell}$ is the observation time closest to t. The approximated continuous time covariate trajectory then changes value half way between each pair of consecutive observation times. That is,

$$z^i_\mu(t) = \sum_{\ell=1}^{n_i} I(t \in [U_{i\ell}, U_{i\ell+1}]) z^i_\mu(t_{i\ell}) \tag{S12}$$

where $U_{i\ell}$ denote the 'switch' times with $U_{i1} = 0$ (the first observation time), $U_{in_i+1} = s_i$ (the final observation time), and all other $U_{i\ell}$ occur half way between consecutive observation times, i.e. $U_{i\ell} = \frac{1}{2}(t_{i\ell-1} + t_{i\ell})$ $\ell = 2, \ldots, n_i$. Using Equation (S12) along with the parameterisations of the association kernels, we can evaluate the integral in Equation (S7) analytically. We do this in the following sections for retarded kernel models A and B.

S1.3.1 Model A. We recall from Equation (16) in the main paper that the association kernel for model A is defined as

$$\beta_\mu(t, t', s) = \frac{a_\mu}{\tau_\mu} \frac{\exp(t'/\tau_\mu)}{\exp(\min(s,t)/\tau_\mu) - 1}. \tag{S13}$$

Therefore, using the step function defined by $\theta(z > 0) = 1$ and $\theta(z < 0) = 0$, we have

$$\begin{aligned}
\mathcal{I}^{(\mathrm{A})}_{ij}[\{a_\mu, \tau_\mu\}] &= \sum_{\mu=1}^{p} \sum_{\ell=1}^{n_j} \frac{a_\mu z^j_\mu(t_{j\ell})}{\mathrm{e}^{\min(s_j,T_i)/\tau_\mu} - 1} \int_0^{\min(s_j,T_i)} \frac{1}{\tau_\mu} \mathrm{e}^{t'/\tau_\mu} I(t' \in [U_{j\ell}, U_{j\ell+1}]) \mathrm{d}t' \\
&= \sum_{\mu=1}^{p} \sum_{\ell=1}^{n_i} \frac{a_\mu z^j_\mu(t_{j\ell})}{\mathrm{e}^{\min(s_j,T_i)/\tau_\mu} - 1} \theta(\min(s_j, T_i, U_{j\ell+1}) - U_{j\ell}) \left[\mathrm{e}^{t'/\tau_\mu} \right]_{U_{j\ell}}^{\min(s_j,T_i,U_{j\ell+1})} \\
&= \sum_{\mu=1}^{p} \sum_{\ell=1}^{n_j} a_\mu z^j_\mu(t_{j\ell}) \, \theta(\min(s_j, T_i, U_{j\ell+1}) - U_{j\ell}) \frac{\mathrm{e}^{\min(s_j,T_i,U_{j\ell+1})/\tau_\mu} - \mathrm{e}^{U_{j\ell}/\tau_\mu}}{\mathrm{e}^{\min(s_j,T_i)/\tau_\mu} - 1} \\
&= \sum_{\mu=1}^{p} \sum_{\ell=1}^{n_j} a_\mu z^j_\mu(t_{j\ell}) \, \theta(T_i - U_{j\ell}) \frac{\mathrm{e}^{\min(T_i,U_{j\ell+1})/\tau_\mu} - \mathrm{e}^{U_{j\ell}/\tau_\mu}}{\mathrm{e}^{\min(s_j,T_i)/\tau_\mu} - 1}
\end{aligned} \tag{S14}$$

where in the last line we used the fact that $U_{j\ell} < U_{j\ell+1} \leq s_j$.

S1.3.2 Model B. We recall from Equation (17) in the main paper that the association kernel for model B is defined as

$$\beta_\mu(t,t',s) = \frac{a_\mu}{\tau_\mu}\mathrm{e}^{-(t-t')/\tau_\mu} + \frac{a_\mu}{\min(s,t)}\left[1-\mathrm{e}^{[\min(s,t)-t]/\tau_\mu}+\mathrm{e}^{-t/\tau_\mu}\right]. \quad \text{(S15)}$$

Substituting this into Equation (S7) gives

$$\begin{aligned}
\mathcal{I}_{ij}^{(\mathrm{B})}[\{a_\mu,\tau_\mu\}] &= \sum_{\mu=1}^{p}\sum_{\ell=1}^{n_j} a_\mu z_\mu^j(t_\ell)\mathrm{e}^{-T_i/\tau_\mu}\int_0^{\min(s_j,T_i)} I(t'\in[U_{j\ell},U_{j\ell+1}])\frac{1}{\tau_\mu}\mathrm{e}^{t'/\tau_\mu}\mathrm{d}t' \\
&\quad + \sum_{\mu=1}^{p}\sum_{\ell=1}^{n_j} a_\mu z_\mu^j(t_\ell)\left(\int_0^{\min(s_j,T_i)} I(t'\in[U_{j\ell},U_{j\ell+1}])\mathrm{d}t'\right)\left(\frac{\mathrm{e}^{-T_i/\tau_\mu}}{\min(s_j,T_i)}+\theta(T_i-s_j)\frac{1-\mathrm{e}^{(s_j-T_i)/\tau_\mu}}{s_j}\right) \\
&= \sum_{\mu=1}^{p}\sum_{\ell=1}^{n_j} a_\mu z_\mu^j(t_\ell)\theta(\min(s_j,T_i,U_{j,\ell+1})-U_{j\ell})\left\{\mathrm{e}^{-T_i/\tau_\mu}\left(\mathrm{e}^{\min(s_j,T_i,U_{j\ell+1})/\tau_\mu}-\mathrm{e}^{U_{j\ell}/\tau_\mu}\right)\right. \\
&\qquad \left. +\left(\min(s_j,T_i,U_{j\ell+1})-U_{j\ell}\right)\left(\frac{\mathrm{e}^{-T_i/\tau_\mu}}{\min(s_j,T_i)}+\theta(T_i-s_j)\frac{1-\mathrm{e}^{(s_j-T_i)/\tau_\mu}}{s_j}\right)\right\} \\
&= \sum_{\mu=1}^{p}\sum_{\ell=1}^{n_j} a_\mu z_\mu^j(t_\ell)\theta(T_i-U_{j\ell})\left\{\mathrm{e}^{-T_i/\tau_\mu}\left(\mathrm{e}^{\min(T_i,U_{j\ell+1})/\tau_\mu}-\mathrm{e}^{U_{j\ell}/\tau_\mu}\right)\right. \\
&\qquad \left. +\left(\min(T_i,U_{j\ell+1})-U_{j\ell}\right)\left(\frac{\mathrm{e}^{-T_i/\tau_\mu}}{\min(s_j,T_i)}+\theta(T_i-s_j)\frac{1-\mathrm{e}^{(s_j-T_i)/\tau_\mu}}{s_j}\right)\right\}, \quad \text{(S16)}
\end{aligned}$$

where in the last line, we have again used the property $U_{j\ell} < U_{j\ell+1} \leq s_j$.

S2 R code for joint models

All joint models were fitted using the `R` package `JMbayes`.[2] All the data analysed in the main paper are available in this package:

1. PBC data
   (a) `pbc2` contains the PBC data set with time varying measurements of covariates
   (b) `pbc2.id` contains the PBC data set with only baseline covariate measurements per individual
2. AIDS data
   (a) `aids` contains the AIDS data set with time varying measurements of covariates
   (b) `aids.id` contains the AIDS data set with only baseline covariate measurements per individual
3. Liver data
   (a) `prothro` contains the Liver data set with time varying measurements of covariates
   (b) `prothros` contains the Liver data set with only baseline covariate measurements per individual.

The models fitted below are specified for the full data sets listed above. For the results presented in the main paper, the data sets were split randomly 20 times into training and test data sets and the models were actually fitted to the training data at each iteration.

S2.1 PBC data

For the results in the main paper we treat the transplant event as a censoring event. To describe this we define a variable `status2` using

```
pbc2.id$status2 <- as.numeric(pbc2.id$status == "dead")
pbc2$status2 <- as.numeric(pbc2$status == "dead")
```

where `status2` $= 1$ if the individual's event is death and $= 0$ otherwise.

For the composite event (results shown in Section S3) we replace `status2` with `status3`,

```
pbc2.id$status3 <- as.numeric(pbc2.id$status != "alive")
pbc2$status3 <- as.numeric(pbc2$status != "alive")
```

defined as 1 if the individual experiences an event (death or a liver transplant) and 0 otherwise (still alive by end of study).

S2.1.1 Linear longitudinal model. Extract of R code used to fit the PBC data set using the simple linear model described in Section 4.2.1 in the main paper. Based on code in Rizopoulos (2012)[3] and Rizopoulos (2018):[4]

```
long.pbc.linear<-mvglmer(list(log(serBilir)~year+(year|id),
                              log(albumin)~year+(year|id),
                              log(prothrombin)~year+(year|id)),
                         data=pbc2, families=list(gaussian, gaussian, gaussian))
surv.pbc<-coxph(Surv(years, status2)~age, data=pbc2.id, model=TRUE)
JM.pbc.linear<-mvJointModelBayes(long.pbc.linear, surv.pbc, timeVar = "year")
```

S2.1.2 Spline model. Extract of R code to fit the PBC data set using the natural cubic spline model described in Section 4.2.1 of the main paper. Based on code in Rizopoulos (2016)[2] and Rizopoulos (2018):[4]

```
long.pbc.spline<-mvglmer(list(log(serBilir)~ns(year,2,B=c(0,14.4))+
                                            (ns(year,2,B=c(0,14.4))|id),
                              log(albumin)~ns(year,2,B=c(0,14.4))+
                                          (ns(year,2,B=c(0,14.4))|id),
                              log(prothrombin)~ns(year,2,B=c(0,14.4))+
                                              (ns(year,2,B=c(0,14.4))|id)),
                         data=pbc2, families=list(gaussian, gaussian, gaussian))
surv.pbc<-coxph(Surv(years, status2)~age, data=pbc2.id, model=TRUE)
JM.pbc.spline<-mvJointModelBayes(long.pbc.spline, surv.spline,
                                 timeVar = "year")
```

S2.2 AIDS data

Extract of R code to fit the AIDS data set using the model described in Section 4.3.1 of the main paper. Based on code in Section 4.2 of Rizopoulos (2012):[3]

```
long.aids<-lme(CD4~obstime+obstime:drug, random=~obstime|patient, data=aids)
surv.aids<-coxph(Surv(Time, death)~drug+prevOI+AZT+gender, data=aids.id,
                 x=TRUE)
JM.aids<-jointModelBayes(long.aids, surv.aids, timeVar="obstime")
```

S2.3 Liver data

Extract of R code to fit the Liver data set using the model described in Section 4.4.1 of the main paper. Replicated from code in Section 5.1.2 of Rizopoulos (2012):[3]

```
prothro$t0<-as.numeric(prothro$time==0)
long.proth<-lme(pro~treat*(ns(time, 3) + t0),
                random=list(id=pdDiag(form=~ns(time,3))), data = prothro)
surv.proth<-coxph(Surv(Time, death)~treat, data=prothros, x=TRUE)
JM.proth<-jointModelBayes(long.proth, surv.proth, timeVar="time")
```

S3 PBC data with composite event

In the main paper we present the results for the PBC data set for models that treat death as the event of interest and transplant events as censoring events. Here we show the results for models that treat the two events (death or transplant) as a single composite event. Figure S1 shows the result for a fixed base time $t = 3$ years and varying prediction time u. Figure S2 shows the results for three fixed prediction windows and varying base time t. With comparison to Figures 5 and 6 in the main paper, we see that the relative accuracy between the models in the two analyses are similar (though overall prediction error for all models is slightly higher for the composite event analysis).

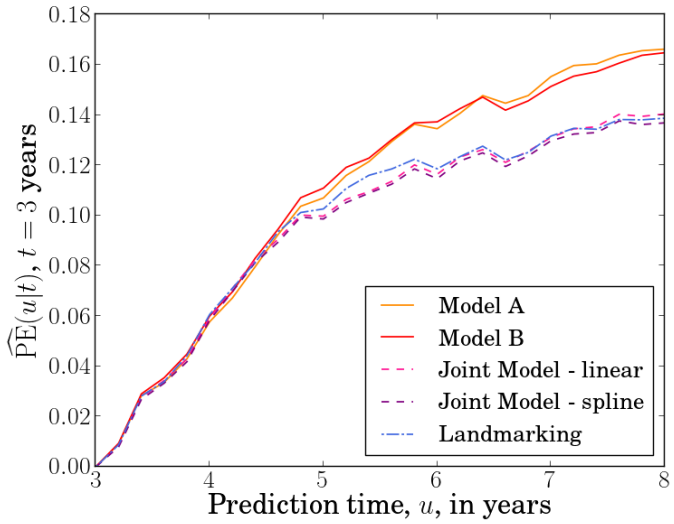

Figure S1. Fixed base time results for the PBC data with models fitted treating the two events (death and transplant) as a single composite event. The plot shows overall prediction error $\widehat{\mathrm{PE}}(u|t)$ as a function of prediction time u (in years) with fixed base time $t = 3$ years. Prediction error is calculated for u values from 3 to 8 years, with 0.2 year increments. A squared loss function was used in Equation (26) in the main paper. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from models A and B of the retarded kernel approach are plotted alongside the landmarking model and two joint models (one that uses a linear longitudinal model for the time-dependent covariates, and another that uses cubic splines). Other than the definition of the composite event, the models fitted are the same as those described in the main paper.

S4 Edits made to prederrJM

The definition of prediction error $\widehat{\mathrm{PE}}(u|t)$ is given in Equation (26) of the main paper. This is identical to the equation for prediction error quoted on pg. 34 in Rizopoulos (2016).[2] For retarded kernel models A and B prediction error is calculated using a `C++` code that exactly follows this equation.

The `JMbayes` package provides the function `prederrJM` to calculate prediction error for joint models (as described in Rizopoulos (2016)[2]). The function can also be used for standard Cox models and, therefore, landmarking models. However, the source code for `prederrJM` varies very slightly from Equation (26). Specifically,

1. `prederrJM` uses $\sum_{i;T_i>t}$ instead of the $\sum_{i;T_i\geq t}$ in Equation (26),
2. for the first term (indiviudals who are still alive), `prederrJM` specifies the condition $I(T_i > u)$ instead of $I(T_i \geq u)$,
3. and for the second term (individuals who have experienced the event), `prederrJM` specifies $\delta_i I(T_i \leq u)$ instead of $\delta_i I(T_i < u)$.

These inconsistencies only have an effect when u or t are exactly equal to one (or more) of the event times T_i in the test data. In the PBC and Liver data sets, event times T_i are quoted to a large number of decimal places meaning we never encounter $u = T_i$ or $t = T_i$ (since we vary t and u in steps of 0.2). However, for the AIDS data set, event times are stored to a lower number of decimal places and we do encounter $u = T_i$ or $t = T_i$ for some values of t and u. For the joint model and landmarking results presented in the main paper we use an edited version of `prederrJM` where inequalities exactly match Equation (26) (and hence the equation for prediction error in Rizopoulos (2016)[2]). This code can be found at the GitHub repository `https://github.com/AnnieDavies/Supplement_Davies_Coolen_Galla_2021`. For the PBC and Liver data sets the results in the main paper are the same as those using the `prederrJM` code without these changed inequalities. Figures S3 and S4 show the results for the AIDS data without these changes. Comparing these to Figures 7 and 8 in the main paper, it is clear the effect of these changes is very minor.

The handling of exceptions in `prederrJM` is such that the function generates an output `NA` if no-one experiences a (non-censoring) event in the window $[t, u]$. Because we are splitting the data sets randomly into training and test sets at different iterations, we occasionally encounter this scenario for certain windows. For the PBC data this occurred for the fixed base time ($t = 3$ years) analysis at prediction time $u = 3.2$ years for iterations 8, 15 and 16, and for the window $w_1 = 1$ year analysis at base time $t = 7.2$ for iteration 20, $t = 7.4$ for iterations 9 and 20, and $t = 7.6, 7.8$ for iteration 16. For the AIDS data (with code edited to match Equation

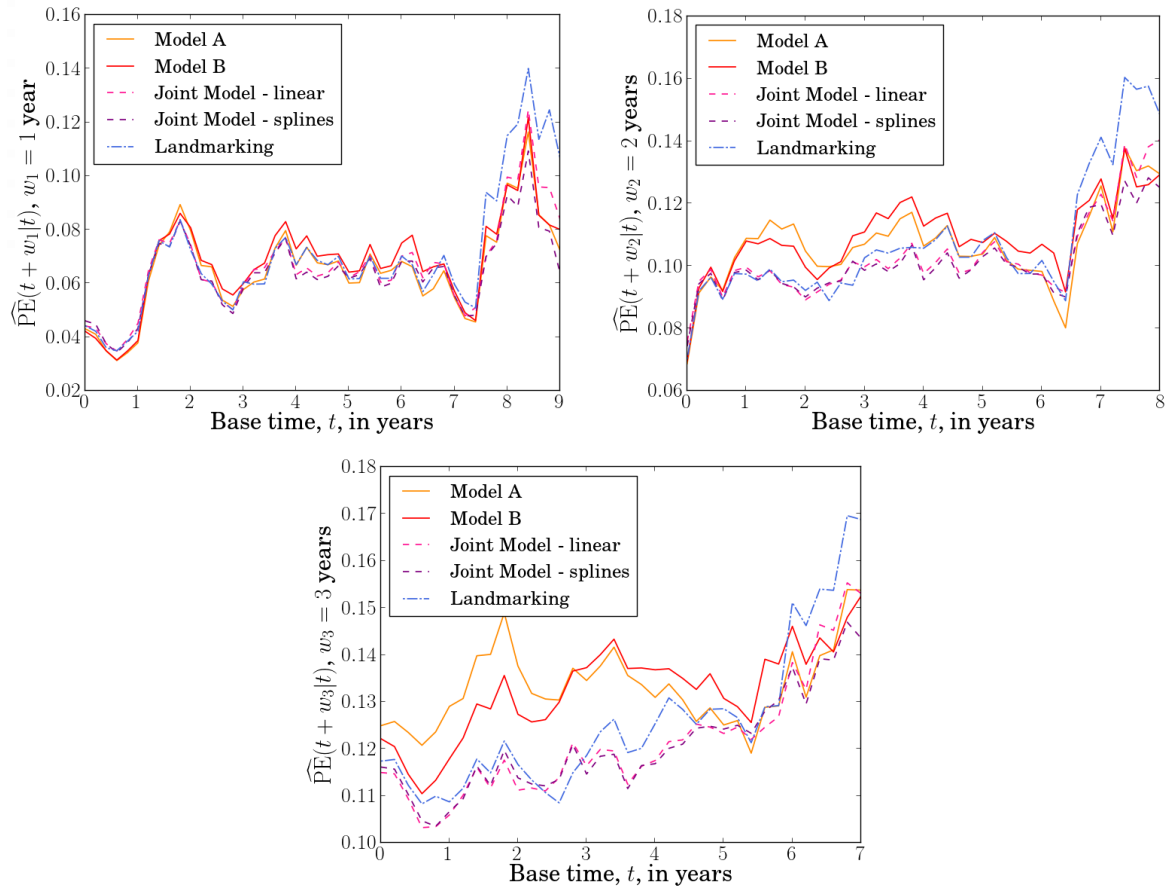

Figure S2. Fixed prediction window results for the PBC data with models fitted treating the two events (death and transplant) as a single composite event. Plots show overall prediction error $\widehat{\mathrm{PE}}(u|t)$ versus base time t (in years), with prediction windows $w_1 = 1$ year, $w_2 = 2$ years and $w_3 = 3$ years. The prediction times are $u = t + w$. The prediction error is calculated for t ranging from 0 to 9,8 or 7 years for w_1, w_2 and w_3 respectively, with 0.2 year increments. A squared loss function was used in Equation (26) in the main paper. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. Results from models A and B of the retarded kernel approach are plotted alongside the landmarking model and two joint models; one that uses a linear longitudinal model for the time-dependent covariates, and another that uses cubic splines. Other than the definition of the composite event, the models fitted are the same as those described in the main paper.

(26)) this occurred only in the fixed base time ($t = 6$ months) analysis at $u = 6.2$ months for iterations 17 and 18 and at $u = 6.4, 6.6$ months for iteration 18. For the Liver data this only occurred for window $w_1 = 1$ year at $t = 8.6, 8.8$ for iteration 16 and $t = 9$ for iteration 5. If there are no non-censoring events in a given window $[t, u]$, the second term in Equation (26) is equal to zero. Therefore, we edited the `prederrJM` source code to handle this scenario (see the Github repository https://github.com/AnnieDavies/Supplement_Davies_Coolen_Galla_2021). The results in the main paper are for this edited code. Compared to the original `prederrJM` code, these edits have a negligible effect on results.

For the version of `prederrJM` for Cox models, we also obtain an output of `NA` if there is no-one censored in the interval $[t, u]$. In the joint model version of `prederrJM` this is handled by including the argument `na.rm=TRUE` when we perform the sum $\sum_{i;T_i \geq t}$. We therefore added this argument to the function for Cox models.

All changes made to the `prederrJM` source code were very minor and had an almost negligible effect on all results. Changes were made to the code only to ensure that all models were evaluated with exactly the same prediction error equation consistent with the equation quoted in literature.[2,5,6]

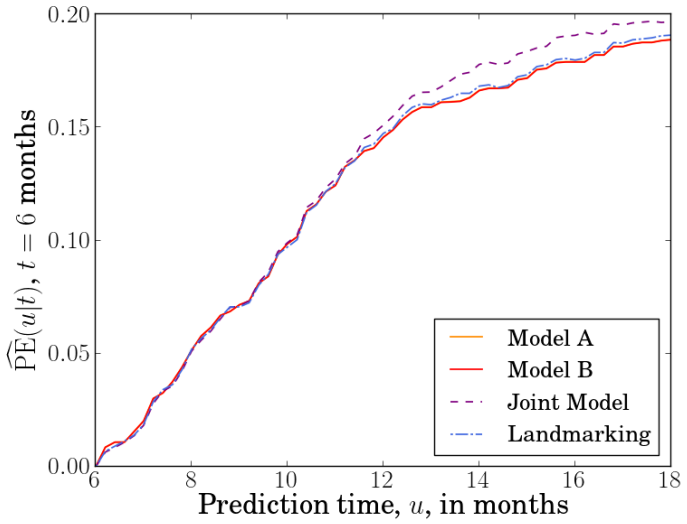

Figure S3. Fixed base time results for the AIDS data set using the `prederrJM` code (for the joint model and landmarking model) without changes made to the inequalities. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ plotted versus prediction time u (in months) for the AIDS data with fixed base time $t = 6$ months. This error is calculated for u ranging from 6 to 18 months, at 0.2 month intervals. In Equation (26) in the main paper a squared loss function was used. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the results from the landmarking model and a joint model. The results from model A (orange line) cannot be seen because they overlap with the results from model B (red line).

S5 Results with decaying association parameter at s=0

The association kernels $\beta_\mu(t, t', s)$ for models A and B as specified in Equations (16) and (17) of the main paper do not hold for $s = 0$. In the data sets we analyse, some individuals are observed only once meaning their final observation time is $s = 0$. For the results presented in the main paper we treat the association parameter of these individuals as fixed, $\beta_\mu(t) = a_\mu$. Another option is to define a decaying parameter, $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$. The results for this latter choice are shown in Figures S5 and S6 for the PBC data (treating transplants as a censoring event), in Figures S7 and S8 for the AIDS data, and in Figures S9 and S10 for the Liver data. For the PBC and Liver data, the results with $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$ and $\beta_\mu(t) = a_\mu$ are similar when the base time $t \gtrsim 1$ year. When we restrict the individuals in the test data to having observations over a smaller time frame, the prediction error for these models is much larger. This effect is increased for the larger prediction windows. This can be understood because for smaller t many individuals in the test data will have been observed only once and the parameter $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$ means the effect of this observation is decayed at later times. Similarly in the AIDS data set, the results are similar to the results in the main paper except when the individuals are restricted to only one observation (at $t = 0$). Perhaps another reasonable choice of association parameter for $s = 0$ is a hybrid of the fixed and decaying association, e.g. $\beta_\mu(t) = a_\mu(1 + \mathrm{e}^{-t/\tau_\mu})$.

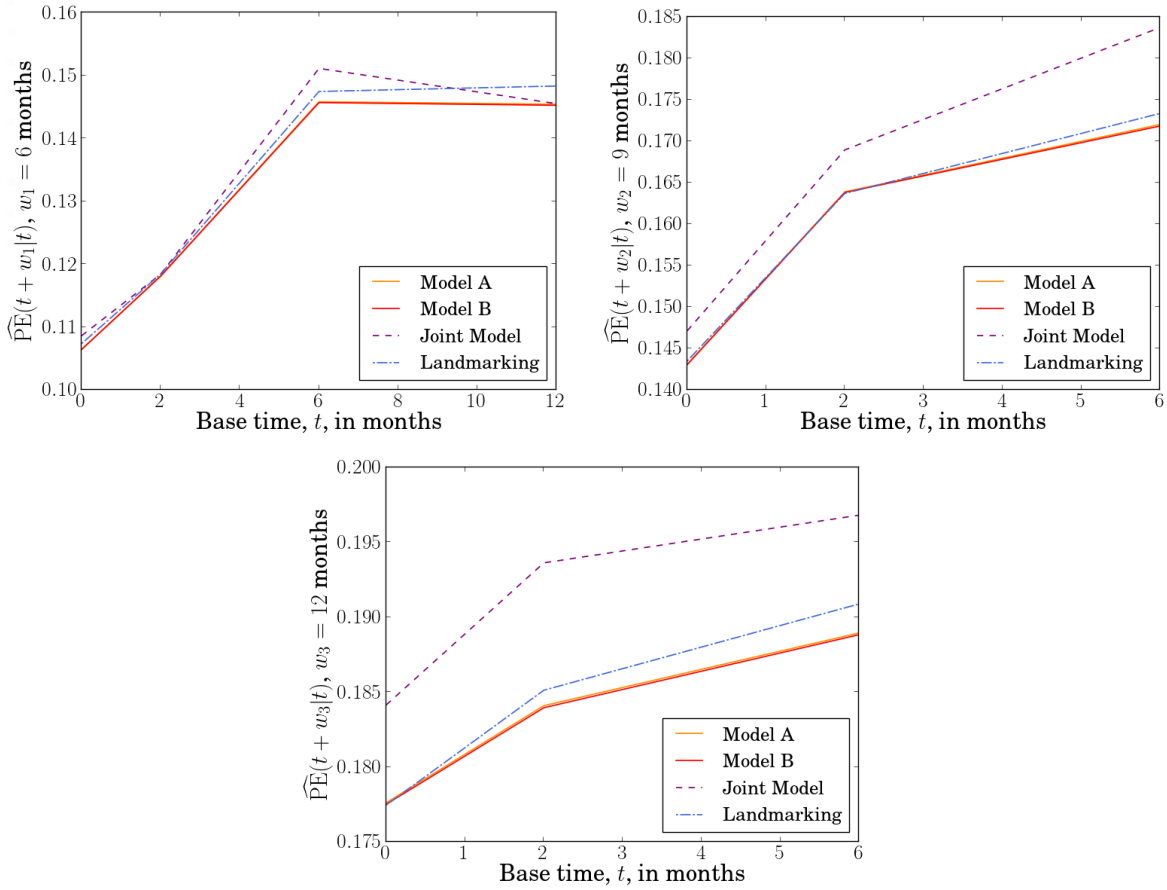

Figure S4. Fixed prediction window results for the AIDS data set using the `prederrJM` code (for the joint model and landmarking model) without changes made to the inequalities. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ versus base time t (in months) for the AIDS data with three fixed prediction windows: $w_1 = 6$ months, $w_2 = 9$ months and $w_3 = 12$ months. The prediction times are $u = t + w$. Observations are made at times $0, 2, 6, 12, 18$ months for all individuals in this data set. Prediction errors are hence only updated at these time points. For prediction window w_1, prediction error is measured for $t = 0, 2, 6$ and 12 months. For windows w_2 and w_3, the error is measured at $t = 0, 2$ and 6 months only. In Equation (26) in the main paper we used a squared loss function. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the landmarking model and a joint model. The results from model A (orange line) cannot be seen clearly because they overlap with the results from model B (red line).

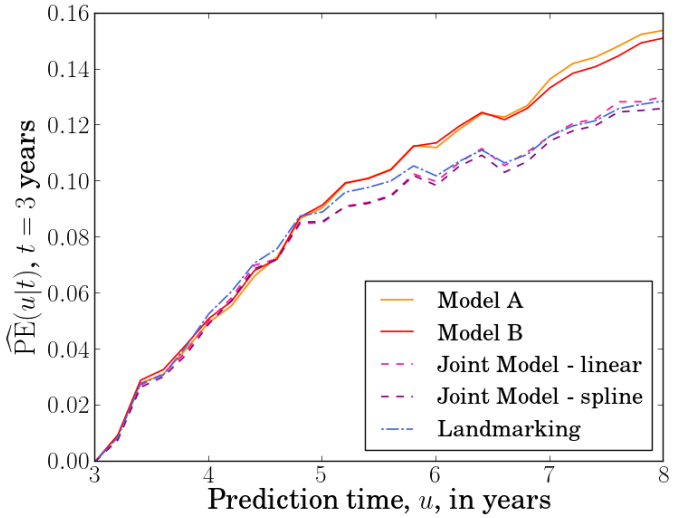

Figure S5. Fixed base time results for the PBC data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. The plot shows overall prediction error $\widehat{\mathrm{PE}}(u|t)$ as a function of prediction time u (in years) with fixed base time $t = 3$ years. Prediction error is calculated for u values from 3 to 8 years, with 0.2 year increments. A squared loss function was used in Equation (26) in the main paper. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from models A and B of the retarded kernel approach are plotted alongside the landmarking model and two joint models (one that uses a linear longitudinal model for the time-dependent covariates, and another that uses cubic splines). Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper (i.e. we treat transplant events as a censoring event).

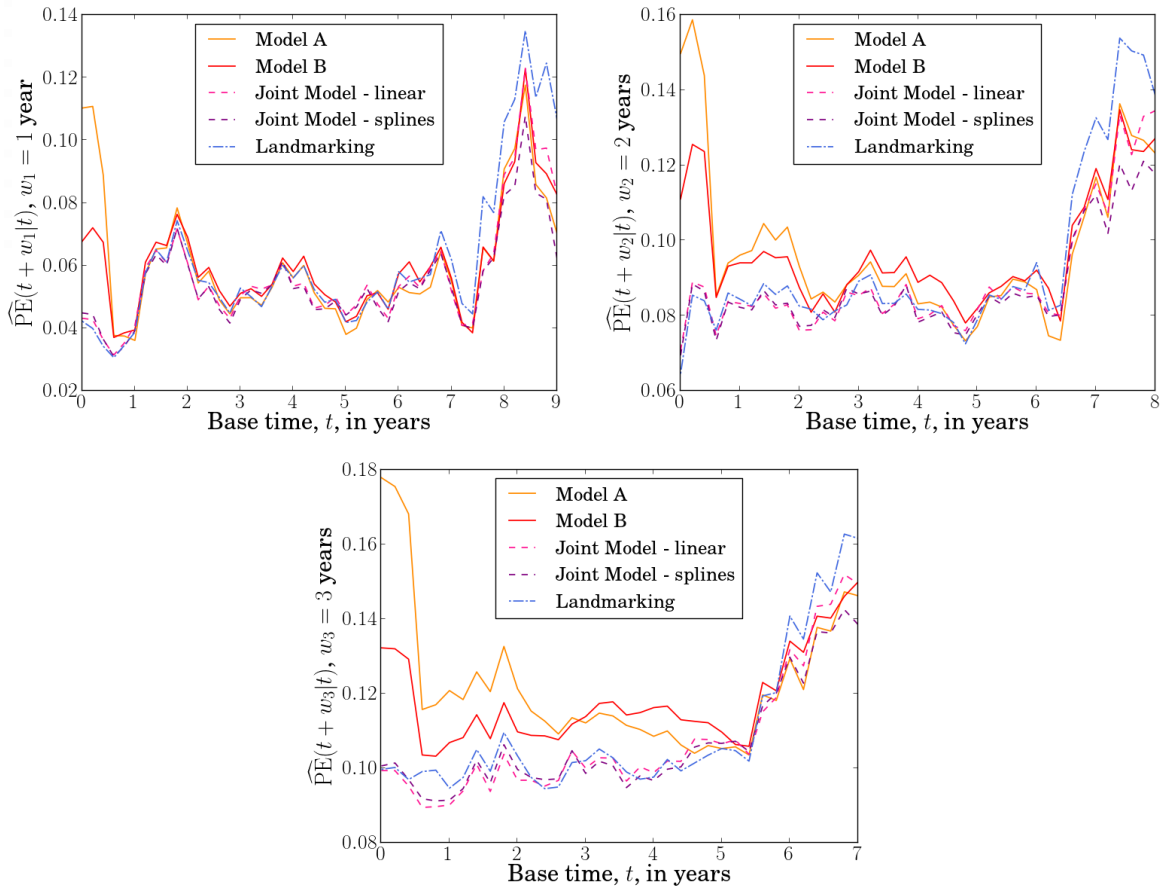

Figure S6. Fixed prediction window results for the PBC data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. Plots show overall prediction error $\widehat{\mathrm{PE}}(u|t)$ versus base time t (in years), with prediction windows $w_1 = 1$ year, $w_2 = 2$ years and $w_3 = 3$ years. The prediction times are $u = t + w$. The prediction error is calculated for t ranging from 0 to 9,8 or 7 years for w_1, w_2 and w_3 respectively, with 0.2 year increments. A squared loss function was used in Equation (26) in the main paper. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. Results from models A and B of the retarded kernel approach are plotted alongside the landmarking model and two joint models; one that uses a linear longitudinal model for the time-dependent covariates, and another that uses cubic splines. Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper (i.e. we treat transplant events as a censoring event).

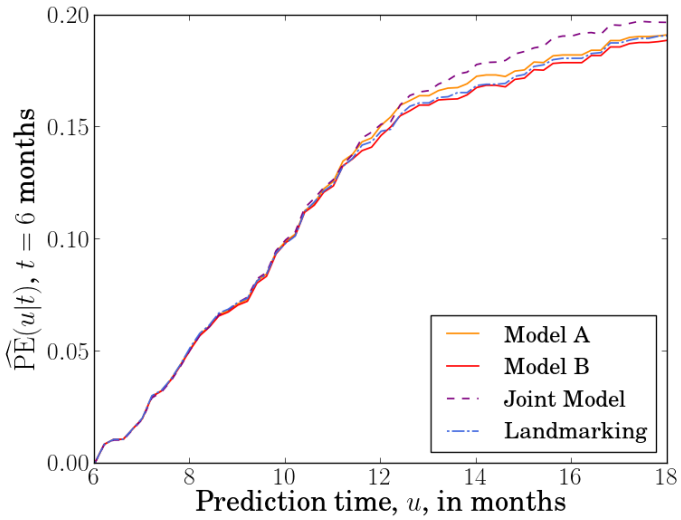

Figure S7. Fixed base time results for the AIDS data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ plotted versus prediction time u (in months) with fixed base time $t = 6$ months. This error is calculated for u ranging from 6 to 18 months, at 0.2 month intervals. In Equation (26) in the main paper a squared loss function was used. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the results from the landmarking model and a joint model. Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper. The results from model A (orange line) cannot be seen because they overlap with the results from model B (red line).

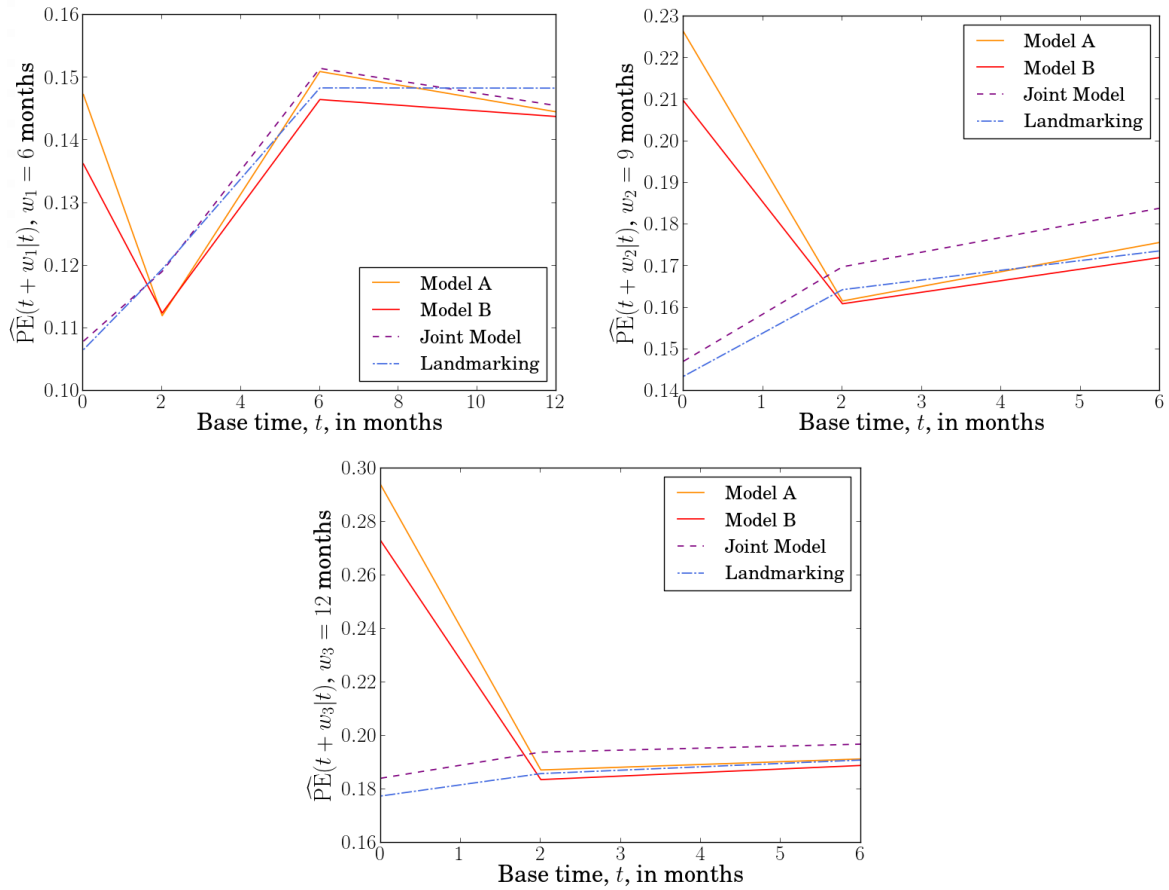

Figure S8. Fixed prediction window results for the AIDS data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ versus base time t (in months) with three fixed prediction windows: $w_1 = 6$ months, $w_2 = 9$ months and $w_3 = 12$ months. The prediction times are $u = t + w$. Observations are made at times $0, 2, 6, 12, 18$ months for all individuals in this data set. Prediction errors are hence only updated at these time points. For prediction window w_1, prediction error is measured for $t = 0, 2, 6$ and 12 months. For windows w_2 and w_3, the error is measured at $t = 0, 2$ and 6 months only. In Equation (26) in the main paper we used a squared loss function. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the landmarking model and a joint model. Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper. The results from model A (orange line) cannot be seen clearly because they overlap with the results from model B (red line).

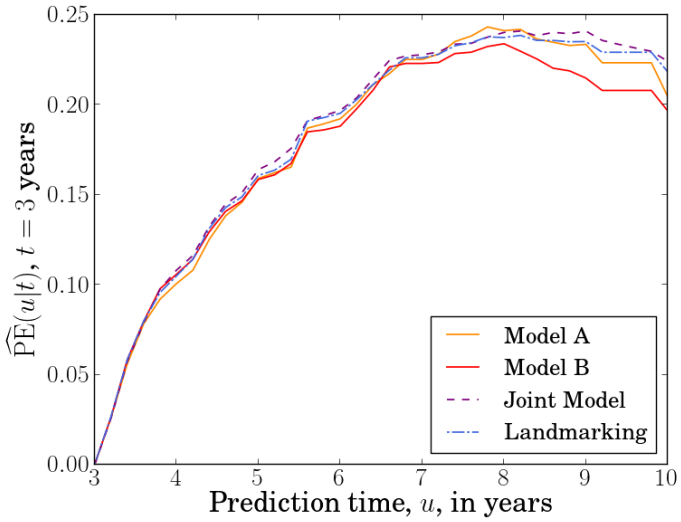

Figure S9. Fixed base time results for the Liver data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ plotted versus prediction time u (in years) with fixed base time $t = 3$ years. This error is calculated for u ranging from 3 to 10 years, with 0.2 year increments. In Equation (26) in the main paper we used a squared loss function. The prediction error plotted at each time u is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the results from the landmarking model and a joint model. Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper.

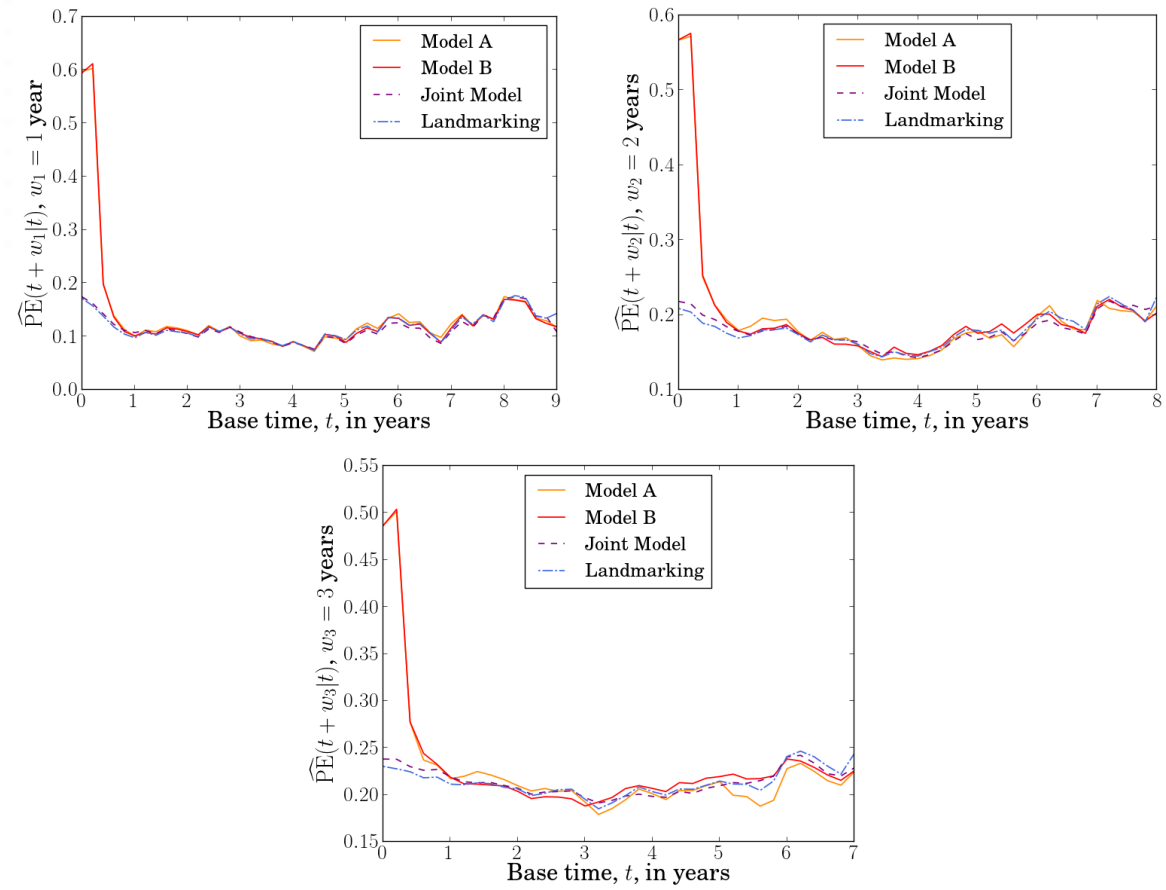

Figure S10. Fixed prediction window results for the Liver data with a decaying association in models A and B, $\beta_\mu(t) = a_\mu \mathrm{e}^{-t/\tau_\mu}$, for individuals who have their final observation time $s = 0$. Overall prediction error $\widehat{\mathrm{PE}}(u|t)$ plotted against base time t (in years) for the Liver data with three fixed prediction windows, $w_1 = 1$ year, $w_2 = 2$ years and $w_3 = 3$ years. The prediction times are $u = t + w$. The error is calculated for t ranging from 0 to 9,8 or 7 years, for w_1, w_2 and w_3 respectively, with 0.2 year intervals. In Equation (26) in the main paper a squared loss function was used. The prediction error plotted at each time t is an average over values of $\widehat{\mathrm{PE}}(u|t)$ calculated for 20 random splits of the data into training and test data sets. The results from retarded kernel models A and B are plotted alongside the landmarking model and a joint model. Other than the definition of the association for $s = 0$ in models A and B, the models fitted are the same as those described in the main paper.

References

1. Breslow NE. Discussion on Professor Cox's paper. *J Roy Stat Soc B Met* 1972; 34: 216–217.
2. Rizopoulos D. The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *J Stat Softw* 2016; 72(7): 1–46.
3. Rizopoulos D. *Joint models for longitudinal and time-to-event data with applications in R*. CRC Biostatistics Series, New York: Chapman and Hall, 2012.
4. Rizopoulos D. Multivariate joint models vignette. Online, 2018. Accessed 07/07/21 from http://www.drizopoulos.com/vignettes/multivariate%20joint%20models.
5. Henderson R, Diggle P and Dobson A. Identification and efficacy of longitudinal markers for survival. *Biostatistics* 2002; 3(1): 33–50.
6. Rizopoulos D, Molenberghs G and Lesaffre EMEH. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical J* 2017; 59(6): 1261–1276.